Robust Audio Scene Analysis for Rescue Robots(Abstract_要旨)

AUTHOR(S):

Bando, Yoshiaki

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	坂東宜昭
論文題目	Robust Audio Scene Analysis for Rescue Robots （レスキューロボットのための頑健な音環境理解）		

（論文内容の要旨）

Audio scene analysis and its application to robot audition is indispensable for a rescue robot, both to detect victims and to sense the robot itself (i.e., location and posture) in disaster environments where visual and/or GPS sensors cannot be used. This work focuses on hose-shaped rescue robots used for probing narrow gaps under rubble. Arrays of microphones, inertial sensors, and loudspeakers on the robot are used for audio scene analysis.

Two fundamental problems of audio scene analysis are addressed: speech enhancement and posture (shape) estimation. Speech enhancement is crucial for detecting speech sounds in captured noisy signals. Posture estimation is essential for enabling an operator to control the flexible robot and for localizing the speech source by using the deformable microphone array. In addition, the integration of these techniques enables the robot to find and approach a victim autonomously. The major difficulties are that the layout of the microphones changes as the robot moves and that some of them are occasionally occluded by rubble. In this study, Bayesian signal processing is used for speech enhancement. The latent speech signals and parameters, which depend on the surrounding environment, are simultaneously estimated . Multi-modal signal processing is used for posture estimation. Unreliable audio measurements due to occlusion are compensated for by using the measurements of other sensors.

This thesis consists of seven chapters. Chapter 2 overviews audio scene analysis for rescue robots and reviews speech enhancement and posture estimation.

Chapter 3 describes a speech enhancement method called Bayesian robust non-negative tensor factorization. To deal with the dynamic configuration of the microphones, speech and noise signals are separated on the basis of their spectral pattern difference instead of on the phase difference (which is unreliable). Under the assumption that the speech and noise spectrograms are sparse and low-rank, respectively, they are separated from the input multichannel spectrogram without prior training. To cope with the partial occlusion of microphones, the speech volume gain at each microphone is estimated on the basis of its feasible gain. Experimental results showed that this method outperforms conventional multichannel methods even when a half of the microphones are occluded.

To further improve the enhancement performance, a deep prior distribution on speech signals is introduced in Chapter 4. Instead of using the unrealistic sparse

assumption for speech signals, a deep generative model is trained with clean speech signals from a large database. Posterior estimates of clean speech are obtained using the speech model as a prior distribution while adapting a noise model to the observed noisy signals. Experimental results showed that this method outperforms a method based on the low-rank and sparse decomposition. The results also showed that the method outperforms a conventional supervised method with deep learning in unseen noisy environments.

Chapter 5 describes an audio-based posture estimation method that can deal with the dynamic configuration of microphones. The time differences of arrival (TDOAs) of beacon sounds, which depend on the locations of the microphones and loudspeakers, are used to estimate the posture. A state-space model representing the posture dynamics is formulated, and the current posture is tracked by estimating the posture change rate and predicting the current posture.

In Chapter 6, the audio-based posture estimation method is extended to a multi-modal 3D posture estimation method that can work when the microphones are partially occluded. The method can automatically exclude TDOA measurements distorted by obstacles and compensate for the missing posture information by using the tilt angles obtained from accelerometers. Experiments using a 3-meter hose-shaped rescue robot showed that this method reduces the tip position error of the initial state to about 0.2 m. When the error of the initial state is less than 20%, it can estimate the correct 3D posture in real time.

Chapter 7 concludes this thesis with a brief look at future work.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３，０００字を標準とすること。
論文内容の要旨を英語で記入する場合は、４００～１，１００wordsで作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること。

（続紙 ２ ）

（論文審査の結果の要旨）

災害現場で被災者を捜索するためのレスキューロボットには、瓦礫などの苛酷な環境下で周辺環境と自己状態をセンシングする機能が求められる。本論文では、マイクロフォンアレイを装着したホース型ロボットを想定し、頑健な音環境理解、特に音声強調と自己姿勢の推定に取り組んだ研究をまとめたもので、主な成果は以下の通りである。

1. 振幅スペクトルにおけるモータ音などの雑音の低ランク性と人間の音声のスパース性に着目し、多チャネル入力信号を用いて、これらを効果的に分離するロバスト非負値テンソル分解(RNTF)という手法を提案した。このモデルを変分ベイズ法により教師なしで推定する方法を定式化・実装し、一部のマイクロフォンの遮蔽があっても頑健に音声強調できることを示した。

2. 低ランク・スパース分解に基づく音声強調の枠組みにおいて、変分オートエンコーダ(VAE)によって音声データベースから事前に深層学習した音声モデルを使う手法(VAE-NMF)を提案した。単チャネル条件において、従来の教師なし手法より高い性能を実現し、教師あり学習による手法と比べても未知の雑音下で頑健に動作することを示した。

3. ホース型ロボットに、マイクロフォンに加えてスピーカも複数塔載し、音の到達時間差の情報からロボットの姿勢とその変化速度を、無香カルマンフィルタ(UKF)を用いて推定する方法を提案し、ロボットが駆動中でもオンラインで姿勢を頑健に推定できることを示した。さらに、加速度センサの情報も統合することで、障害物がある条件下でも３次元の姿勢を頑健に推定できることを示した。

以上のように本論文は、ベイズ学習などの統計モデルに基づいて、マイクロフォンの移動と遮蔽を許容する頑健な音声強調と姿勢推定を実現する方法を提示したもので、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成30年 2月15日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降